WORKING PAPERS

N° 1366

Revised version December 2024

Learning Markov Processes with Latent Variables

Koen Jochmans, Ayden Higgins

Toulouse School of Economics

LEARNING MARKOV PROCESSES WITH LATENT VARIABLES*

Ayden Higgins

University of Exeter Business School
Rennes Drive
Exeter EX4 4PU
United Kingdom

a.higgins@exeter.ac.uk

Koen Jochmans

Toulouse School of Economics
Université Toulouse Capitole
1 esplanade de l'Université
31080 Toulouse CEDEX 6
France

koen.jochmans@tse-fr.eu

This version: December 4, 2024

*We are grateful to Anna Mikusheva, Peter Phillips, and two anonymous referees for constructive feedback.
Financial support from the European Research Council, grant ERC-2016-STG-715787, and from the French Government and the ANR under the Investissements d' Avenir program, grant ANR-17-EURE-0010 is gratefully acknowledged.

Proposed running head: Markov processes with latent variables

Corresponding author: Koen Jochmans

Abstract: We consider the problem of identifying the parameters of a time-homogeneous bivariate Markov chain when only one of the two variables is observable. We show that, subject to conditions that we spell out, the transition kernel and the distribution of the initial condition are uniquely recoverable (up to an arbitrary relabelling of the state space of the latent variable) from the joint distribution of four (or more) consecutive time-series observations. The result is, therefore, applicable to (short) panel data as well as to (stationary) time series data.

JEL Classification: C32, C33, C38

Keywords: dynamic discrete choice, finite mixture, Markov process, regime switching, state dependence.

1 Introduction

Let $Y_0, Y_1, \ldots, Y_T$ be a bivariate random process on a finite state space. Once the random variable Y_0 has been drawn from an initial distribution, the sequence $Y_1, \ldots, Y_T$ evolves according to a time-homogeneous Markov chain. Partition Y_t as (X_t, Z_t). The random variables $X_0, X_1, \ldots, X_T$ can take on r values and are observable. The random variables $Z_0, Z_1, \ldots, Z_T$ can take on q values and are latent. We complete the model with the assumption that the transition probability

$$\mathbb{P}(X_t = x', Z_t = z' | X_{t-1} = x, Z_{t-1} = z)$$

factors as

$$\mathbb{P}(X_t = x' | X_{t-1} = x, Z_t = z') \times \mathbb{P}(Z_t = z' | X_{t-1} = x, Z_{t-1} = z). \tag{1}$$

This is a redundancy statement on further lags of the latent variable and is intuitive. Our aim is to recover the distribution of the initial condition Y_0 and the (time-invariant) transition probabilities from Y_{t-1} to Y_t from the distribution of $X_0, X_1, \ldots, X_T$ alone. We show that, subject to conditions that are spelled out below, this is possible as soon as three (consecutive) transitions are observed, i.e., $T \geq 3$. Here, identification is to be understood as being up to an arbitrary relabeling of the state space of the Z_t. Such a permutational ambiguity is standard in models with latent variables and is harmless for our current purposes.

The state space of the X_t is taken to be the set of positive integers up to r and the state space of the Z_t is normalized to the set of positive integers up to q. The former restriction

is imposed for notational convenience—translation to a general set is immediate—while the latter normalization, given that the process in question is unobserved, is without loss of generality. Also, given that the state spaces are finite, restricting attention to scalar random variables is innocuous.[1]

While we focus on identification, our results justify estimation from short panel data with non-stationary initial conditions as well as from a single long time series, provided that the initial condition is drawn from the steady state distribution. In either case, given the discreteness of the variables involved, estimation can be done by maximum likelihood, typically using a version of the EM algorithm.[2] This is well understood; see, for example, Ailliot and Pène (2015) and Pouzo, Psaradakis and Sola (2022) for details. Procedures to recover parameters of structural dynamic discrete-choice models that fit our framework are

[1]Suppose that we observe k-dimensional vectors $\boldsymbol{X}_t$ whose entries can take on, respectively, $r_1, \ldots, r_k$ values. Enumerate all values in the state space of $\boldsymbol{X}_t$ and define a scalar random variable X_t on this set of numbers as a known one-to-one transformation of $\boldsymbol{X}_t$. Such a construction is always possible. The state space of the induced X_t consists of $r = r_1 \times \cdots \times r_k$ values. Identification of the Markov process $Y_t = (X_t, Z_t)$ then implies identification of the corresponding process $(\boldsymbol{X}_t, Z_t)$. Similarly, if we are interested in a scalar outcome in the presence of a (possibly vector valued) discrete covariate, we may define $\boldsymbol{X}_t$ to be the vector containing both the outcome of interest and the covariate and proceed to identify the Markovian dynamics of $(\boldsymbol{X}_t, Z_t)$ in the manner just described. From this identification of the conditional law then follows readily.

[2]Alternatively, as our identification argument is constructive, an estimator based on it can be developed. Computationally this can be achieved by working with the algorithm put forth in Higgins and Jochmans (2021). Asymptotics for such an estimator could be derived by following arguments along the lines of those in Bonhomme, Jochmans and Robin (2016a).

given in Hotz and Miller (1993), Bajari, Benkard and Levin (2007), Arcidiacono and Miller (2011), and Connault (2016), among others.

Versions of the model that we study have been used in applied work, an early example is Miller (1984), and the question of identification has been investigated by Hu and Shum (2012).[3] While their result has a flavor that is similar to ours, their approach and the assumptions underlying it are different in several respects. We detail these differences below.

The setup under study here also encompasses the hidden Markov model (see, e.g., Cappé, Moulines and Rydén 2005, Gassiat, Cleynen and Robin 2016, and Bonhomme, Jochmans and Robin 2016a) and the multivariate mixture model (Anderson 1954, Hall and Zhou 2003, Hu 2008, Kasahara and Shimotsu 2009, Bonhomme, Jochmans and Robin 2016a,b, Vandermeulen and Scott 2020, Higgins and Jochmans 2023). Even though our arguments bear some similarity with some of the approaches to identification taken there, the fact that the observed and unobserved variables are allowed to be jointly Markovian makes the key restrictions used there inapplicable here.

For the main part of the paper we will be concerned with the question of identification from knowledge of the joint distribution of four time-series observations. In Section 2 we set

[3] They look at the case where $r = q$, also allowing for the variables X_t and Z_t to both be continuous. In addition, Hu and Shum (2012) also consider the situation where the transition kernel is time-varying, in which case they obtain identification results only for the transitions between a subset of the available time periods. Although we do not focus on this here, the arguments we develop in this paper could equally be used in that extended setup.

up notation for what is directly identified from data and for the primitive parameters of the model. In Section 3 we state the assumptions under which we will work, provide discussion, and state our main result. In Section 4 we then present a detailed proof. In Section 5 we discuss how our assumptions and approach differ from Hu and Shum (2012). In Section 6 we show how our approach can be adjusted when additional time-series observations are available. In Section 7 we conclude.

2 Observables and model primitives

We will work with decompositions of the joint distribution of X_0, X_1, X_2, X_3 and subsets thereof. These distributions only involve observable variables and may, thus, be considered known for our purposes. It will be convenient to collect the various probability distributions in the form of a set of matrices. First, the two-way table of (X_0, X_1) is given by the $r \times r$ matrix

$$(\boldsymbol{P})_{x_1,x_0} \coloneqq \mathbb{P}(X_1 = x_1, X_0 = x_0).$$

Next, the three-way table of (X_0, X_1, X_2) is contained in the set of r matrices

$$(\boldsymbol{P}_{x_1})_{x_2,x_0} \coloneqq \mathbb{P}(X_2 = x_2, X_1 = x_1, X_0 = x_0),$$

indexed by $1 \leq x_1 \leq r$, each of which is $r \times r$. Finally, the distribution of all observable variables (X_0, X_1, X_2, X_3) is collected in the set of r^2 matrices of size $r \times r$, indexed by $1 \leq x_1 \leq r$ and $1 \leq x_2 \leq r$,

$$(\boldsymbol{P}_{x_2,x_1})_{x_3,x_0} \coloneqq \mathbb{P}(X_3 = x_3, X_2 = x_2, X_1 = x_1, X_0 = x_0).$$

Throughout we use x_0, x_1, x_2, x_3 to denote particular values that X_0, X_1, X_2, X_3 may take. We note that, because (X_0, Z_0) need not be drawn from any steady-state distribution, the time-series process is not stationary, in general, and so, on occasion, it is important to be explicit on how certain matrices depend on the different time periods. We also recall that the state space of the observable variables is the set of integers up to r so that the values x_0, x_1, x_2, x_3 all range across that set; they are, thus, indices that can be used to indicate matrix entries.

The primitive parameters that we aim to identify are the $q \times r$ matrix of initial conditions

$$(\boldsymbol{\Omega})_{z_0,x_0} \coloneqq \mathbb{P}(Z_0 = z_0, X_0 = x_0)$$

and the collection of r^2 matrices of size $q \times q$, doubly-indexed by $1 \leq x \leq r$ and $1 \leq x' \leq r$,

$$(\boldsymbol{\Theta}_{x',x})_{z',z} \coloneqq \mathbb{P}(X_t = x', Z_t = z' | X_{t-1} = x, Z_{t-1} = z),$$

that, together, make up the transition kernel of $Y_t = (X_t, Z_t)$. Given these parameters, the distribution of $Y_t = (X_t, Z_t)$ for any $t > 0$ can be recovered by starting at the distribution of the initial condition and iterating on the transition kernel.

Because the Z_t are unobservable the elements of their state space, already normalized to be $1, \ldots, q$, can be permuted without any observable implications. This is an ambiguity that is inherent in our specification that can only be resolved by assigning empirical content to the latent variables. For our purposes it does imply that we can only hope to recover $\boldsymbol{\Omega}$ and the $\boldsymbol{\Theta}_{x',x}$ for $1 \leq x \leq r$ and $1 \leq x' \leq r$ up to suitable re-arrangement. Moreover, we

will aim to learn

$$\boldsymbol{\Delta}^{-1}\boldsymbol{\Omega} \qquad \text{and} \qquad \boldsymbol{\Delta}^{-1}\boldsymbol{\Theta}_{x',x}\boldsymbol{\Delta}$$

for $1 \leq x \leq r$ and $1 \leq x' \leq r$, where $\boldsymbol{\Delta}$ is an arbitrary (but common) permutation matrix.

3 Assumptions and main result

Our identification argument makes use of a set of three assumptions, which we provide next.

Our first assumption involves the matrices $\boldsymbol{P}_x$, $1 \leq x \leq r$, which are directly observable. To interpret the restrictions it is, nonetheless, useful to connect them to the primitive parameters of the model. To do so we do need to introduce some additional notation. First consider the $r \times q$ matrices

$$(\boldsymbol{\Xi}_x)_{x',z} \coloneqq \mathbb{P}(X_t = x' | X_{t-1} = x, Z_t = z)$$

and the $q \times q$ matrices

$$(\boldsymbol{\Sigma}_x)_{z',z} \coloneqq \mathbb{P}(Z_t = z' | X_{t-1} = x, Z_{t-1} = z)$$

where, in each case, $1 \leq x \leq r$. All of these matrices are time-invariant. From (1) they may be seen to contain the components of the transition kernel of our Markov process. Furthermore, the product $\boldsymbol{\Xi}_x\boldsymbol{\Sigma}_x$ marginalizes with respect to the latent state Z_t, yielding

$$(\boldsymbol{\Xi}_x\boldsymbol{\Sigma}_x)_{x',z} = \mathbb{P}(X_t = x' | X_{t-1} = x, Z_{t-1} = z). \tag{2}$$

Next we introduce two other sets of matrices, again for $1 \leq x_1 \leq r$, but on this occasion time-dependent. These are, first, the $r \times q$ matrices

$$(\boldsymbol{\Phi}_{x_1})_{x_0,z_1} \coloneqq \mathbb{P}(X_0 = x_0 | X_1 = x_1, Z_1 = z_1), \tag{3}$$

and, second, the $q \times q$ diagonal matrices

$$(\boldsymbol{\Pi}_{x_1})_{z_1,z_1} \coloneqq \mathbb{P}(X_1 = x_1, Z_1 = z_1). \tag{4}$$

The matrix $\boldsymbol{\Phi}_{x_1}$ is a transition matrix for the time-reversed process. The matrices $\boldsymbol{\Pi}_{x_1}$ for $1 \leq x_1 \leq r$, in turn, contain the joint distribution of $Y_1 = (X_1, Z_1)$. We remark that $(\boldsymbol{\Phi}_{x_1}\boldsymbol{\Pi}_{x_1})_{x_0,z_1} = \mathbb{P}(X_0 = x_0, X_1 = x_1, Z_1 = z_1)$. A small calculation reveals that we have the factorization

$$\boldsymbol{P}_{x_1} = (\boldsymbol{\Xi}_{x_1}\boldsymbol{\Sigma}_{x_1})\,\boldsymbol{\Pi}_{x_1}\,\boldsymbol{\Phi}_{x_1}^{\top} \tag{5}$$

for $1 \leq x_1 \leq r$. This decomposition is a consequence of the fact that X_0 and X_2 are independent conditional on (X_1, Z_1).

Our first assumption reads as follows.

Assumption 1. *For each $1 \leq x_1 \leq r$, (i) the $r \times r$ matrix $\boldsymbol{P}_{x_1}$ has rank q or, equivalently, (ii) the columns of the $r \times q$ matrices $\boldsymbol{\Xi}_{x_1}$ and $\boldsymbol{\Phi}_{x_1}$ are linearly independent, and the $q \times q$ matrices $\boldsymbol{\Sigma}_{x_1}$ and $\boldsymbol{\Pi}_{x_1}$ are invertible.*

The formulation in (i) is convenient as it reveals that Assumption 1 is easily testable from the data by means of any of a number of procedures; one example is the rank test of Kleibergen and Paap (2006).

The formulation in (ii) is useful to see what is needed from the underlying Markov process. The decomposition in (5) is similar to factorizations encountered in the analysis of multivariate mixtures (Hall and Zhou 2003; Bonhomme, Jochmans and Robin 2016a), with conditional distributions corresponding to the columns of the $r \times q$ matrices $\boldsymbol{\Xi}_{x_1}\boldsymbol{\Sigma}_{x_1}$ and $\boldsymbol{\Phi}_{x_1}$, and mixing distribution $\boldsymbol{\Pi}_{x_1}$.

Assumption 1 demands that the time-invariant matrices $\boldsymbol{\Xi}_x\boldsymbol{\Sigma}_x$ for $1 \leq x \leq r$ have full column rank, meaning that changes in the latent state at period $t-1$ have sufficient implication on the observable state in period t. This, in turn, can be seen to require that the transition matrix of the latent state, $\boldsymbol{\Sigma}_x$, is invertible for all $1 \leq x \leq r$ and that the conditional distributions of the observed state—i.e., the columns of $\boldsymbol{\Xi}_x$—are linearly independent for all $1 \leq x \leq r$. These conditions are familiar from the literature on hidden Markov models (Gassiat, Cleynen and Robin 2016; Bonhomme, Jochmans and Robin 2016a). At the same time, Assumption 1 imposes that $\boldsymbol{\Phi}_{x_1}\boldsymbol{\Pi}_{x_1}$ for $1 \leq x_1 \leq r$ have maximal column rank. This can be interpreted as a restriction on the initial condition; the variable (X_1, Z_1) needs to have full support over the state space $\{1, \ldots, r\} \times \{1, \ldots, q\}$, and changes in Z_1 need to have a sufficient impact on the conditional distribution of X_0 given (X_1, Z_1). Taken together, the conditions in Assumption 1(ii) ensure that the mixture factorization in (5) is irreducibe, that is, that it cannot be written as a mixture of fewer than q components. The need for this is intuitive, and it is a standard requirement in the analysis of multivariate latent-variable models (Hall and Zhou 2003, Hu 2008, Allman, Matias and Rhodes 2009, Kasahara and Shimotsu 2009, Hu and Shum 2012, Bonhomme,

Jochmans and Robin 2016a,b, Vandermeulen and Scott 2020, Higgins and Jochmans 2023).

Our remaining assumptions further restrict the matrices $\boldsymbol{\Theta}_{x',x}$. To motivate why they are needed we note that, under Assumption 1 we can (as we will show in the proof below), for each $1 \leq x \leq r$, construct $r \times q$ matrices $\boldsymbol{U}_x$ and $\boldsymbol{V}_x$ from $\boldsymbol{P}_x$ for which we have that

$$\dot{\boldsymbol{P}}_{x',x} \coloneqq \boldsymbol{U}_{x',x}^{\top} \boldsymbol{P}_{x',x} \boldsymbol{V}_x = \boldsymbol{\mathcal{Q}}_{x'} \boldsymbol{\Theta}_{x',x} \boldsymbol{\mathcal{Q}}_x^{-1} \tag{6}$$

for an unobserved $q \times q$ matrix $\boldsymbol{\mathcal{Q}}_x$ and for all $1 \leq x \leq r$ and $1 \leq x' \leq r$. This suggests the possibility to recover the components of the transition kernel by solving the above equation,

$$\boldsymbol{\mathcal{Q}}_{x'}^{-1} \dot{\boldsymbol{P}}_{x',x} \boldsymbol{\mathcal{Q}}_x = \boldsymbol{\Theta}_{x',x}.$$

As such, the task of learning the transition kernel of our Markov process has been recast into identifying the collection of matrices $\boldsymbol{\mathcal{Q}}_x$ for $1 \leq x \leq r$. Under Assumptions 2 and 3 this can be achieved up to a common permutation of their columns. That is, we may recover

$$\boldsymbol{Q}_x \coloneqq \boldsymbol{\mathcal{Q}}_x \boldsymbol{\Delta}, \qquad 1 \leq x \leq r,$$

for some $q \times q$ permutation matrix $\boldsymbol{\Delta}$ that does not depend on x. With the transition kernel recovered, identification of the distribution of the initial condition, the only other primitive parameter, will follow readily.

Assumption 2 will permit us to characterize the $\boldsymbol{\mathcal{Q}}_x$ for $1 \leq x \leq r$ as eigenvectors of certain matrices, and to recover $\boldsymbol{\mathcal{Q}}_x \boldsymbol{\Delta}_x$ for $1 \leq x \leq r$, where the $\boldsymbol{\Delta}_x$ are unknown permutation matrices that, in general, will depend on the value x in question in an arbitrary manner. Assumption 3, then, will allow us to obtain an ordering that is independent of

x, and thus to learn $\boldsymbol{Q}_x$ for $1 \leq x \leq r$. The remaining permutational ambiguity is due to the inherent invariance of an eigendecomposition to the ordering of eigenvalues and eigenvectors and, in the current exercise, reflects the fact that the state space of the latent Z_t can be relabelled without observable implications.

To facilitate the exposition we state Assumption 2 in two parts.

Assumption 2. *(i) For each* $1 \leq x' \leq r$ *there exists at least one* $1 \leq x \leq r$ *so that* $\mathbb{P}(X_t = x' | X_{t-1} = x, Z_t = z) > 0$ *for all* $1 \leq z \leq q$*. Let* $\mathcal{X}_{x'}$ *be the set of values* x *for which this holds for a given* x'*. (ii) For each* $1 \leq x' \leq r$ *there exists at least one* $1 \leq x'' \leq r$ *so that* $\mathcal{X}_{x'} \neq \mathcal{X}_{x''}$*.*

As will become apparent on inspection of the proof below, (i) is verifiable for a given pair (x', x) by checking that the $q \times q$ matrix $\dot{\boldsymbol{P}}_{x',x}$ is invertible. Again, this may be done by using a standard rank test. Similarly, (ii) is testable via a collection of such tests. Together, (i) and (ii) ensure that, for each $1 \leq x \leq r$, matrix $\boldsymbol{\mathcal{Q}}_x$ can be cast as a matrix of eigenvectors.

Assumption 2 (Continued). *(iii) For each* $1 \leq x' \leq r$ *and for all pairs* $1 \leq z < z' \leq q$ *there exist an* x'' *as in (ii) so that, for some* $\dot{x} \in \mathcal{X}_{x'}$ *and* $\ddot{x} \in \mathcal{X}_{x''}$ *with* $\dot{x} \neq \ddot{x}$*, it holds that*

$$\frac{\mathbb{P}(X_t = x'' | X_{t-1} = \dot{x}, Z_t = z)}{\mathbb{P}(X_t = x' | X_{t-1} = \dot{x}, Z_t = z)} \frac{\mathbb{P}(X_t = x' | X_{t-1} = \ddot{x}, Z_t = z)}{\mathbb{P}(X_t = x'' | X_{t-1} = \ddot{x}, Z_t = z)}$$

is different from

$$\frac{\mathbb{P}(X_t = x'' | X_{t-1} = \dot{x}, Z_t = z')}{\mathbb{P}(X_t = x' | X_{t-1} = \dot{x}, Z_t = z')} \frac{\mathbb{P}(X_t = x' | X_{t-1} = \ddot{x}, Z_t = z')}{\mathbb{P}(X_t = x'' | X_{t-1} = \ddot{x}, Z_t = z')}.$$

Part (iii) is a necessary and sufficient condition for $\boldsymbol{\mathcal{Q}}_x$, $1 \leq x \leq r$, to be unique up to permutation and scaling of their columns (De Lathauwer, De Moor and Vandewalle 2004, Theorem 6.1).

Below we show that the matrix $\boldsymbol{P}$ contains sufficient information to resolve the scaling indeterminacy. To be able to recover the $\boldsymbol{\mathcal{Q}}_x$ for $1 \leq x \leq r$ up to a common (across $1 \leq x \leq r$) column permutation, however, we need an additional restriction. We use the following.

Assumption 3. *There exists a value $1 \leq \dot{x} \leq r$ such that for each $1 \leq x' \leq r$ we can find an $1 \leq x \leq r$ (which may change with x') so that $\mathbb{P}(X_t = x' | X_{t-1} = x, Z_t = z) > 0$ and $\mathbb{P}(X_t = \dot{x} | X_{t-1} = x, Z_t = z) > 0$ both hold for all $1 \leq z \leq q$.*

Together, Assumptions 2 and 3 restrict the distributions that make up the columns of $\boldsymbol{\Xi}_x$ for $1 \leq x \leq r$ beyond what is imposed by the linear-independence requirement in Assumption 1. For example, Assumption 2(i) demands that for each $1 \leq x' \leq r$ there is a matrix $\boldsymbol{\Xi}_x$ for which row x' contains no zeros. Similarly, Assumption 3 needs there to exist an integer $\dot{x}$ so that for each $1 \leq x' \leq r$ there is a matrix $\boldsymbol{\Xi}_x$ whose $\dot{x}$th and x'th row contain no zeros.

In the next section we prove Theorem 1.

Theorem 1. *Let Assumptions 1–3 hold. Then the distribution of $Y_0 = (X_0, Z_0)$ and the transition matrix of the Markov process $Y_t = (X_t, Z_t)$ are identified from the joint distribution of X_0, X_1, X_2, X_3, up to an arbitrary relabelling of the state space of the latent Z_t.*

4 Proof of Theorem 1

We proceed in three steps. First we use the model restrictions together with Assumption 1 to arrive at (6). Next, we use (1) and Assumptions 2-3 to recover the $\boldsymbol{Q}_x$ for $1 \leq x \leq r$. Finally, with these matrices in hand, we complete the argument by identifying our primitive parameters.

Step 1 (Multilinear restrictions). By Assumption 1, for each $1 \leq x \leq r$, the $r \times r$ matrix $\boldsymbol{P}_x$ has rank q. Hence, it admits the singular-value decomposition $\boldsymbol{P}_x = \bar{\boldsymbol{U}}_x \boldsymbol{S}_x \bar{\boldsymbol{V}}_x^\top$ for $r \times q$ matrices $\bar{\boldsymbol{U}}_x$ and $\bar{\boldsymbol{V}}_x$ that satisfy $\bar{\boldsymbol{U}}_x^\top \bar{\boldsymbol{U}}_x = \boldsymbol{I}_q$ and $\bar{\boldsymbol{V}}_x^\top \bar{\boldsymbol{V}}_x = \boldsymbol{I}_q$, and invertible $q \times q$ diagonal matrix $\boldsymbol{S}_x$. Here, and later, $\boldsymbol{I}_q$ is the $q \times q$ identity matrix. It follows that, with $\boldsymbol{U}_x \coloneqq \boldsymbol{S}_x^{-1/2} \bar{\boldsymbol{U}}_x^\top$ and $\boldsymbol{V}_x \coloneqq \boldsymbol{S}_x^{-1/2} \bar{\boldsymbol{V}}_x^\top$, we have $\boldsymbol{U}_x \boldsymbol{P}_x \boldsymbol{V}_x^\top = \boldsymbol{I}_q$. When combined with (5), this implies that

$$\boldsymbol{I}_q = \boldsymbol{U}_x \boldsymbol{P}_x \boldsymbol{V}_x^\top = \boldsymbol{U}_x (\boldsymbol{\Xi}_x \boldsymbol{\Sigma}_x \boldsymbol{\Pi}_x \boldsymbol{\Phi}_x^\top) \boldsymbol{V}_x^\top = (\boldsymbol{U}_x \boldsymbol{\Xi}_x \boldsymbol{\Sigma}_x) (\boldsymbol{V}_x \boldsymbol{\Phi}_x \boldsymbol{\Pi}_x)^\top = \boldsymbol{\mathcal{Q}}_x \boldsymbol{\mathcal{Q}}_x^{-1}$$

where the $q \times q$ matrix $\boldsymbol{\mathcal{Q}}_x \coloneqq (\boldsymbol{U}_x \boldsymbol{\Xi}_x \boldsymbol{\Sigma}_x)$ is full rank by Assumption 1 and the fact that the equality $(\boldsymbol{V}_x \boldsymbol{\Phi}_x \boldsymbol{\Pi}_x)^\top = \boldsymbol{\mathcal{Q}}_x^{-1}$ holds as a consequence of the construction in the above display.

Now, in the same way as (5), we have, for all $1 \leq x \leq r$ and $1 \leq x' \leq r$, the factorization

$$\boldsymbol{P}_{x',x} = (\boldsymbol{\Xi}_{x'} \boldsymbol{\Sigma}_{x'}) \boldsymbol{\Theta}_{x',x} (\boldsymbol{\Pi}_x \boldsymbol{\Phi}_x^\top). \tag{7}$$

Therefore,

$$\boldsymbol{U}_{x'} \boldsymbol{P}_{x',x} \boldsymbol{V}_x^\top = \dot{\boldsymbol{P}}_{x',x} = (\boldsymbol{U}_{x'} \boldsymbol{\Xi}_{x'} \boldsymbol{\Sigma}_{x'}) \boldsymbol{\Theta}_{x',x} (\boldsymbol{V}_x \boldsymbol{\Phi}_x \boldsymbol{\Pi}_x)^\top = \boldsymbol{\mathcal{Q}}_{x'} \boldsymbol{\Theta}_{x',x} \boldsymbol{\mathcal{Q}}_x^{-1},$$

which corresponds to (6).

Step 2 (Matrix diagonalization). Note that, by (1), for all $1 \leq x \leq r$ and $1 \leq x' \leq r$,

$$\boldsymbol{\Theta}_{x',x} = \boldsymbol{\Upsilon}_{x',x}\, \boldsymbol{\Sigma}_x,$$

for the $q \times q$ diagonal matrix $(\boldsymbol{\Upsilon}_{x',x})_{z,z} := (\boldsymbol{\Xi}_x)_{x',z}$ containing the x'th row of $\boldsymbol{\Xi}_x$. Therefore, (6) gives

$$\dot{\boldsymbol{P}}_{x',x} = \boldsymbol{\mathcal{Q}}_{x'} \boldsymbol{\Upsilon}_{x',x}\, \boldsymbol{\Sigma}_x \boldsymbol{\mathcal{Q}}_x^{-1}$$

for all $1 \leq x \leq r$ and $1 \leq x' \leq r$.

Now, fix x' and let $\dot{x}$ be so that Assumption 2(i) holds, i.e., $\dot{x} \in \mathcal{X}_{x'}$. Then $\dot{\boldsymbol{P}}_{x',\dot{x}}$ is invertible and

$$\dot{\boldsymbol{P}}_{x',\dot{x}}^{-1} = \boldsymbol{\mathcal{Q}}_{\dot{x}}\, \boldsymbol{\Sigma}_{\dot{x}}^{-1} \boldsymbol{\Upsilon}_{x',\dot{x}}^{-1} \boldsymbol{\mathcal{Q}}_{x'}^{-1}.$$

Hence, for any $1 \leq x'' \leq r$ we have that

$$\dot{\boldsymbol{P}}_{x'',\dot{x}} \dot{\boldsymbol{P}}_{x',\dot{x}}^{-1} = (\boldsymbol{\mathcal{Q}}_{x''} \boldsymbol{\Upsilon}_{x'',\dot{x}}\, \boldsymbol{\Sigma}_{\dot{x}} \boldsymbol{\mathcal{Q}}_{\dot{x}}^{-1})(\boldsymbol{\mathcal{Q}}_{\dot{x}}\, \boldsymbol{\Sigma}_{\dot{x}}^{-1} \boldsymbol{\Upsilon}_{x',\dot{x}}^{-1} \boldsymbol{\mathcal{Q}}_{x'}^{-1}) = \boldsymbol{\mathcal{Q}}_{x''} \boldsymbol{\Upsilon}_{x'',\dot{x}} \boldsymbol{\Upsilon}_{x',\dot{x}}^{-1} \boldsymbol{\mathcal{Q}}_{x'}^{-1},$$

for $\boldsymbol{\Upsilon}_{x'',\dot{x}} \boldsymbol{\Upsilon}_{x',\dot{x}}^{-1}$ diagonal. Next, let $\ddot{x} \in \mathcal{X}_{x''}$ be different from $\dot{x}$. By Assumption 2(i)–(ii) such a pair $(x'', \ddot{x})$ exists. Then, in the same way as before,

$$\dot{\boldsymbol{P}}_{x',\ddot{x}} \dot{\boldsymbol{P}}_{x'',\ddot{x}}^{-1} = (\boldsymbol{\mathcal{Q}}_{x'} \boldsymbol{\Upsilon}_{x',\ddot{x}}\, \boldsymbol{\Sigma}_{\ddot{x}} \boldsymbol{\mathcal{Q}}_{\ddot{x}}^{-1})(\boldsymbol{\mathcal{Q}}_{\ddot{x}}\, \boldsymbol{\Sigma}_{\ddot{x}}^{-1} \boldsymbol{\Upsilon}_{x'',\ddot{x}}^{-1} \boldsymbol{\mathcal{Q}}_{x''}^{-1}) = \boldsymbol{\mathcal{Q}}_{x'} \boldsymbol{\Upsilon}_{x',\ddot{x}} \boldsymbol{\Upsilon}_{x'',\ddot{x}}^{-1} \boldsymbol{\mathcal{Q}}_{x''}^{-1}$$

is well defined. Multiplying both matrices yields

$$(\dot{\boldsymbol{P}}_{x',\ddot{x}} \dot{\boldsymbol{P}}_{x'',\ddot{x}}^{-1})(\dot{\boldsymbol{P}}_{x'',\dot{x}} \dot{\boldsymbol{P}}_{x',\dot{x}}^{-1}) = \boldsymbol{\mathcal{Q}}_{x'}(\boldsymbol{\Upsilon}_{x',\ddot{x}} \boldsymbol{\Upsilon}_{x'',\ddot{x}}^{-1} \boldsymbol{\Upsilon}_{x'',\dot{x}} \boldsymbol{\Upsilon}_{x',\dot{x}}^{-1}) \boldsymbol{\mathcal{Q}}_{x'}^{-1}; \tag{8}$$

the right-hand side of this equation constitutes an eigendecomposition. The eigenvalues depend on all of $x', x'', \dot{x}, \ddot{x}$ while the eigenvectors depend only on x'. Thus, for each $1 \leq x' \leq r$ there will, in general, be multiple matrices that are diagonalizable in the same basis. Assumption 2(iii) concerns the eigenvalues. Moreover, it ensures that there is sufficient distinctness in them such that the eigenvectors are unique up to scaling and permutation. That is, for a diagonal matrix $\boldsymbol{\Lambda}_x$ and a permutation matrix $\boldsymbol{\Delta}_x$, the matrix

$$\tilde{\boldsymbol{Q}}_x \coloneqq \boldsymbol{\mathcal{Q}}_x \boldsymbol{\Lambda}_x \boldsymbol{\Delta}_x$$

is identified for $1 \leq x \leq r$.

Let $\boldsymbol{p}'_x$ be the xth row of the matrix $\boldsymbol{P}$ and write $\boldsymbol{\iota}_q$ for the q-vector of all ones. Observe that, by (3) and (4), we have $\boldsymbol{p}'_x = \boldsymbol{\iota}_q^\top (\boldsymbol{\Pi}_x \boldsymbol{\Phi}_x^\top)$. Therefore, using the functional form of $\boldsymbol{\mathcal{Q}}_x^{-1}$, we have that $\boldsymbol{p}'_x \boldsymbol{V}_x^\top = \boldsymbol{\iota}_q^\top \boldsymbol{\mathcal{Q}}_x^{-1}$. Thus, right-multiplying both sides of this expression by $\tilde{\boldsymbol{Q}}_x$ we obtain

$$\boldsymbol{p}'_x \boldsymbol{V}_x^\top \tilde{\boldsymbol{Q}}_x = \boldsymbol{\iota}_q^\top \boldsymbol{\Lambda}_x \boldsymbol{\Delta}_x = \boldsymbol{\iota}_q^\top (\boldsymbol{\Delta}_x^{-1} \boldsymbol{\Lambda}_x \boldsymbol{\Delta}_x),$$

where the last equality follows from the fact that $\boldsymbol{\Delta}_x^{-1}$ is a permutation matrix and, thus, has exactly one entry of 1 in each row and each column and all other entries equal to 0. The above equation returns the main diagonal of a diagonal matrix and so equally the matrix

$$\tilde{\boldsymbol{\Lambda}}_x \coloneqq \boldsymbol{\Delta}_x^{-1} \boldsymbol{\Lambda}_x \boldsymbol{\Delta}_x$$

itself. With this matrix in hand, we are then able to construct $\tilde{\boldsymbol{Q}}_x \tilde{\boldsymbol{\Lambda}}_x^{-1} = \boldsymbol{\mathcal{Q}}_x \boldsymbol{\Delta}_x$ for all $1 \leq x \leq r$.

To see how we may recover all matrices up to a common permutation of their columns, consider a value x' and, for this value, let the pair $(x, \dot{x})$ be as in Assumption 3. Then the matrix $\dot{\boldsymbol{P}}_{x',x}\dot{\boldsymbol{P}}_{\dot{x},x}^{-1} = \boldsymbol{\mathcal{Q}}_{x'}\boldsymbol{\Upsilon}_{x',x}\boldsymbol{\Upsilon}_{\dot{x},x}^{-1}\boldsymbol{\mathcal{Q}}_{\dot{x}}^{-1}$ is well-defined. Pre- and post-multiplication with $\tilde{\boldsymbol{\Lambda}}_{x'}\tilde{\boldsymbol{Q}}_{x'}^{-1}$ and $\tilde{\boldsymbol{Q}}_{\dot{x}}\tilde{\boldsymbol{\Lambda}}_{\dot{x}}^{-1}$, respectively, gives

$$(\tilde{\boldsymbol{\Lambda}}_{x'}\tilde{\boldsymbol{Q}}_{x'}^{-1})(\dot{\boldsymbol{P}}_{x',x}\dot{\boldsymbol{P}}_{\dot{x},x}^{-1})(\tilde{\boldsymbol{Q}}_{\dot{x}}\tilde{\boldsymbol{\Lambda}}_{\dot{x}}^{-1}) = \boldsymbol{\Delta}_{x'}^{-1}\boldsymbol{\Upsilon}_{x',x}\boldsymbol{\Upsilon}_{\dot{x},x}^{-1}\boldsymbol{\Delta}_{\dot{x}} = \boldsymbol{\Delta}_{x'}^{-1}\boldsymbol{\Delta}_{\dot{x}}(\boldsymbol{\Delta}_{\dot{x}}^{-1}\boldsymbol{\Upsilon}_{x',x}\boldsymbol{\Upsilon}_{\dot{x},x}^{-1}\boldsymbol{\Delta}_{\dot{x}}).$$

Notice that $\boldsymbol{\Delta}_{x'}^{-1}\boldsymbol{\Delta}_{\dot{x}}$ is a permutation matrix and that $\boldsymbol{\Delta}_{\dot{x}}^{-1}\boldsymbol{\Upsilon}_{x',x}\boldsymbol{\Upsilon}_{\dot{x},x}^{-1}\boldsymbol{\Delta}_{\dot{x}}$ is a diagonal matrix. This latter matrix thus corresponds to the (in general, non-diagonal) matrix on the left-hand side up to a re-ordering of its rows. The columnwise sum of the left-hand side matrix thus yields $\boldsymbol{\Delta}_{\dot{x}}^{-1}\boldsymbol{\Upsilon}_{x',x}\boldsymbol{\Upsilon}_{\dot{x},x}^{-1}\boldsymbol{\Delta}_{\dot{x}}$. Assumption 3 ensures this matrix to be invertible. Therefore, we can solve for

$$\boldsymbol{H}_{x',\dot{x}} \coloneqq \boldsymbol{\Delta}_{x'}^{-1}\boldsymbol{\Delta}_{\dot{x}}.$$

The argument can be applied for each $1 \leq x' \leq r$ using the same $\dot{x}$. Given these matrices we can compute

$$\tilde{\boldsymbol{Q}}_x\,\tilde{\boldsymbol{\Lambda}}_x^{-1}\boldsymbol{H}_{x,\dot{x}} = \boldsymbol{\mathcal{Q}}_x\boldsymbol{\Delta}_{\dot{x}},$$

which corresponds to $\boldsymbol{Q}_x$ for $\boldsymbol{\Delta} = \boldsymbol{\Delta}_{\dot{x}}$ and for all $1 \leq x \leq r$.

Step 3 (Parameter recovery). With $\boldsymbol{Q}_x$ for $1 \leq x \leq r$ in hand, from (6), we readily obtain

$$\check{\boldsymbol{\Theta}}_{x',x} \coloneqq \boldsymbol{Q}_{x'}^{-1}\dot{\boldsymbol{P}}_{x',x}\,\boldsymbol{Q}_x = \boldsymbol{\Delta}^{-1}\boldsymbol{\Theta}_{x',x}\,\boldsymbol{\Delta}$$

for all $1 \leq x \leq r$ and $1 \leq x' \leq r$ which allows to assemble a coherent transition kernel for the Markov process of $Y_t = (X_t, Z_t)$.

Next, let $\boldsymbol{\omega}_x$ be the xth column of $\boldsymbol{\Omega}$ and let $\boldsymbol{p}_x$ be the xth column of the matrix $\boldsymbol{P}$. The two sets of vectors are related through $\boldsymbol{p}_x = (\boldsymbol{\Xi}_x \boldsymbol{\Sigma}_x)\, \boldsymbol{\omega}_x$ for all $1 \leq x \leq r$. Given that the transition kernel has been recovered up to $\boldsymbol{\Delta}$ we also know $\boldsymbol{R}_x \coloneqq (\boldsymbol{\Xi}_x \boldsymbol{\Sigma}_x)\boldsymbol{\Delta}$ for all $1 \leq x \leq r$. By Assumption 1 these matrices all have maximal column rank. We can, therefore, uniquely solve the linear system $\boldsymbol{p}_x = \boldsymbol{R}_x \check{\boldsymbol{\omega}}_x$ for $\check{\boldsymbol{\omega}}_x \coloneqq \boldsymbol{\Delta}^{-1}\boldsymbol{\omega}_x$, yielding identification of

$$\check{\boldsymbol{\omega}}_x \coloneqq (\boldsymbol{R}_x^\top \boldsymbol{R}_x)^{-1} \boldsymbol{R}_x^\top \boldsymbol{p}_x$$

for all $1 \leq x \leq r$. Collecting these vectors in the matrix $\check{\boldsymbol{\Omega}}_x$ identifies $\boldsymbol{\Delta}^{-1}\boldsymbol{\Omega}$. With all primitive parameters identified up to the common permutation $\boldsymbol{\Delta}$, the proof of Theorem 1 is complete.

5 Discussion

In earlier work Hu and Shum (2012) gave an identification result similar to Theorem 1. Their argument, and the assumptions underlying it, are related to ours but differ in several respects.

First, in their version of Assumption 1, Hu and Shum (2012) assume that the matrices $\boldsymbol{P}_x$ for $1 \leq x \leq r$ are invertible. This demands that $r = q$, that is, that the support of X_t and the support of Z_t have the same cardinality. Of course, the case where $r < q$ is outside the scope of Theorem 1 (although it is within the confines of its extension discussed below) but it is clear from our derivations that, all else equal, a larger r cannot make the

identification problem more complicated. As in our case, a key step in the proof of Hu and Shum (2012) is an eigendecomposition. Moreover, similar to (8), one has, when $r = q$,

$$(\boldsymbol{P}_{x',\ddot{x}}\boldsymbol{P}_{x'',\ddot{x}}^{-1})(\boldsymbol{P}_{x'',\dot{x}}\boldsymbol{P}_{x',\dot{x}}^{-1}) = (\boldsymbol{\Xi}_{x'}\boldsymbol{\Sigma}_{x'})(\boldsymbol{\Upsilon}_{x',\ddot{x}}\boldsymbol{\Upsilon}_{x'',\ddot{x}}^{-1}\boldsymbol{\Upsilon}_{x'',\dot{x}}\boldsymbol{\Upsilon}_{x',\dot{x}}^{-1})(\boldsymbol{\Xi}_{x'}\boldsymbol{\Sigma}_{x'})^{-1}, \tag{9}$$

which is well defined under Assumptions 1 and 2. A difference between (8) and (9) that Hu and Shum (2012) exploit is that in the latter the eigenvectors are known to be valid probability mass functions. That is, they are known to sum to one, and so their scale is known.

Hu and Shum (2012) then recover the matrices $(\boldsymbol{\Xi}_{x'}\boldsymbol{\Sigma}_{x'})\boldsymbol{\Delta}_{x'}$ as eigenvectors from (9) for $1 \leq x' \leq r$. To ensure uniqueness (up to the permutation matrix) they impose a version of Assumption 2 that is stronger than needed. Whereas we, in (8), exploit the fact that there are, in general, multiple matrices that are jointly diagonalizable in the same basis, they consider the eigendecomposition in (9) for a single matrix. Although their decomposition is not contained in (8), it is clear that its eigenvalues correspond to those of one of its members. Their Assumption 3 thus deals with the same ratios as does our Assumption 2(iii). However, as they do not consider joint diagonalization, they require that, for each $1 \leq x' \leq r$ in Assumption 2(iii), there exists a triple of values $(x'', \dot{x}, \ddot{x})$ for which all q eigenvalues are different.

For the transition kernel of the full Markov process to be recoverable from their results up to this point one needs to be able to enforce a common ordering on the columns, i.e., find a transformation $\boldsymbol{\Delta}_x^{-1}\boldsymbol{\Delta}$. We used Assumption 3 to do so. Hu and Shum (2012) combine the fact that the columns of $(\boldsymbol{\Xi}_x\boldsymbol{\Sigma}_x)$ are probability mass functions with a monotonicity

condition on one of their functionals to be able to re-arrange them in a common order. Concretely, they assume that, for each $1 \leq x \leq r$, there is a known functional, such as the mean or median, of the distribution $\mathbb{P}(X_t = x' | X_{t-1} = x, Z_{t-1} = z)$ that is strictly monotone in z. The plausibility of such a condition depends on the context at hand. For example, in the multivariate setting discussed in Footnote 1 it would appear difficult to maintain.

When $r > q$, (9) does not go through and the above argument can no longer be applied. While a version of (9) could be constructed after binning the support of X_t into $q < r$ groups, an application of the argument of Hu and Shum (2012) would only allow to recover conditional probability distributions defined over these binned states, and not over the full state space.

6 Generalization

Identification from the joint distribution of $X_0, X_1, \ldots, X_T$ for larger T can be done by a minor adaptation of our arguments. The key insight remains that, for any $0 < t < T$, the vectors $(X_{t+1}, X_{t+2}, \ldots, X_T)$ and $(X_0, X_1, \ldots, X_{t-1})$ are independent conditional on (X_t, Z_t). Let $\lfloor \cdot \rfloor$ be the floor function, so that $\lfloor a \rfloor$ is the greatest integer less than or equal to a. Then we can combine the probabilities $\mathbb{P}(X_{T-1} = x_{T-1}, \ldots, X_{\lfloor T/2 \rfloor} = x_{\lfloor T/2 \rfloor}, \ldots, X_0 = x_0)$ into the collection of matrices

$$\boldsymbol{P}_{x_{\lfloor T/2 \rfloor}}, \qquad 1 \leq x_{\lfloor T/2 \rfloor} \leq r,$$

where the rows vary with the values $(x_{\lfloor T/2\rfloor-1}, x_{\lfloor T/2\rfloor-2} \ldots, x_0)$ and the columns vary with the values $(x_{T-1}, x_{T-2}, \ldots, x_{\lfloor T/2\rfloor+1})$; each such matrix is thus of dimension $r^{T-\lfloor T/2\rfloor-1} \times r^{\lfloor T/2\rfloor}$. Similarly, we can collect the probabilities $\mathbb{P}(X_T = x_T, \ldots, X_{\lfloor T/2\rfloor} = x_{\lfloor T/2\rfloor}, \ldots, X_0 = x_0)$ into the matrices

$$\boldsymbol{P}_{x_{\lfloor T/2\rfloor+1}, x_{\lfloor T/2\rfloor}}, \qquad 1 \leq x_{\lfloor T/2\rfloor} \leq r,\ 1 \leq x_{\lfloor T/2\rfloor+1} \leq r,$$

each of which is again of dimension $r^{T-\lfloor T/2\rfloor-1} \times r^{\lfloor T/2\rfloor}$. These matrices admit a factorization akin to, respectively, (5) and (7), and we may then work through the same steps of the proof in Section 4 to obtain identification. Now that the matrices $\boldsymbol{P}_x$ are larger, the demand that $\operatorname{rank} \boldsymbol{P}_x = q$ becomes less stringent. Hence, Assumption 1 becomes less demanding. Assumptions 2 and 3 are used in the exact same way as before and require no modification.

7 Concluding remarks

Our identification result relies on having access to (the distribution of) three consecutive observations, along with an initial condition. There is some reason to believe that this number cannot be improved upon. In particular, this is known to be so for the special case of multivariate finite mixtures (Vandermeulen and Scott 2020, Theorem 4.5).[4] The fact

[4] In their analysis the demand that the component distributions are linearly independent can be relaxed at the expense of requiring that the number of observations is at least as large as $2q - 1$ (refer to their Theorem 4.1). Teicher (1963) provided an earlier version of this result in a more specific context and, more recently, Alexandrovich, Holzmann and Leister (2016) obtained the equivalent in the setting of the hidden Markov model.

that the observations are consecutive is also important in the development of our argument. Modifying our approach to show identification from, for example, the distribution of three non-consecutive transitions is not immediate.

Still, there may be scope for alternative identifying restrictions when fewer than three consecutive transitions are available. Although we are not aware of any such results at the generality of the model we entertain here, some such alternative conditions have been obtained in special cases. Jochmans, Henry and Salanié (2017) exploited tail restrictions in mixture models to circumvent the partial-identification result of Hall and Zhou (2003) and Henry, Kitamura and Salanié (2014), while Gupta, Kumar and Vassilvitskii (2016) used excess-support requirements to learn dynamic mixture models from observing only two (consecutive) transitions.

References

Ailliot, P. and F. Pène (2015). Consistency of the maximum likelihood estimate for non-homogeneous Markov-switching models. *ESAIM: Probability and Statistics 19*, 268–292.

Alexandrovich, G., H. Holzmann, and A. Leister (2016). Nonparametric identification and maximum likelihood estimation of hidden Markov models. *Biometrika 103*, 423–434.

Allman, E. S., C. Matias, and J. A. Rhodes (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics 37*, 3099–3132.

Anderson, T. W. (1954). On estimation of parameters in latent structure analysis. *Psy-*

chometrika 19, 1–10.

Arcidiacono, P. and R. A. Miller (2011). Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica 79*, 1823–1867.

Bajari, P., L. Benkard, and J. Levin (2007). Estimating dynamic models of imperfect competition. *Econometrica 75*, 1331–1370.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016a). Estimating multivariate latent-structure models. *Annals of Statistics 44*, 540–563.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016b). Non-parametric estimation of finite mixtures from repeated measurements. *Journal of the Royal Statistical Society - Series B 78*, 211–229.

Cappé, O., E. Moulines, and T. Rydén (2005). *Inference in hidden Markov models*. Springer Series in Statistics. Springer.

Connault, B. (2016). Hidden Rust models. Mimeo.

De Lathauwer, L., B. De Moor, and J. Vandewalle (2004). Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition. *SIAM Journal of Matrix Analysis and Applications 26*, 295–327.

Gassiat, E., A. Cleynen, and S. Robin (2016). Finite state space non parametric hidden Markov models are in general identifiable. *Statistics and Computing 26*, 61–71.

Gupta, R., R. Kumar, and S. Vassilvitskii (2016). On mixtures of Markov chains. Thirtieth Conference on Neural Information Processing Systems, Barcelona.

Hall, P. and X.-H. Zhou (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics 31*, 201–224.

Henry, M., Y. Kitamura, and B. Salanié (2014). Partial identification of finite mixtures in econometric models. *Quantitative Economics 5*, 123–144.

Higgins, A. and K. Jochmans (2021). Joint approximate asymmetric diagonalization of non-orthogonal matrices. Mimeo.

Higgins, A. and K. Jochmans (2023). Identification of mixtures of dynamic discrete choices. *Journal of Econometrics 237*, 105462.

Hotz, J. and R. Miller (1993). Conditional choice probabilities and the estimation of dynamic models. *Review of Economic Studies 60*, 497–529.

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics 144*, 27–61.

Hu, Y. and M. Shum (2012). Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics 171*, 32–44.

Jochmans, K., M. Henry, and B. Salanié (2017). Inference on two-component mixtures under tail restrictions. *Econometric Theory 33*, 610–635.

Kasahara, H. and K. Shimotsu (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica 77*, 135–175.

Kleibergen, F. and R. Paap (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics 133*, 97–126.

Miller, R. A. (1984). Job matching and occupational choice. *Journal of Political Economy 71*, 1565–1578.

Pouzo, D., Z. Psaradakis, and M. Sola (2022). Maximum likelihood estimation in Markov regime-switching models with covariate-dependent transition probabilities. *Econometrica 90*, 1681–1710.

Teicher, H. (1963). Identifiability of finite mixtures. *Annals of Mathematical Statistics 34*, 1265–1269.

Vandermeulen, R. A. and C. D. Scott (2020). An operator theoretic approach to nonparametric mixture models. *Annals of Statistics 47*, 2704–2733.